Data Analysis & Machine Learning

Academic Year: 2021-22

Course organiser: Christos Leonidopoulos

School of Physics & Astronomy
University of Edinburgh

March 8, 2022

1 Introduction

Data Analysis and Machine Learning (DAML) is a post-graduate 20-credit course, offered to MSc students in Particle and Nuclear Physics. The course is also open to PhD students in Experimental Particle or Nuclear Physics, and undergraduate students. It includes a review of theoretical concepts and practical lab sessions. The course aims to train students on advanced physics-analysis methods and computational techniques invoked in experimental particle or nuclear physics projects.

The DAML course is *not* a substitute for dedicated programming, statistics or machine-learning courses.

1.1 Lecturers

- Prof. Christos Leonidopoulos (course organiser): `Christos.Leonidopoulos@ed.ac.uk`
- Prof. Peter Clarke: `Peter.Clarke@ed.ac.uk`
- Dr. Guillermo Hamity: `hamity.daml@pm.me`
- Dr. Ben Wynne: `b.m.wynne@ed.ac.uk`

1.2 Teaching Assistants

- Semester-1:
  - Joe Bradley: `s1406647@sms.ed.ac.uk`
  - Nisha Grewal: `N.Grewal@sms.ed.ac.uk`
  - Victoria Parrish: `V.A.Parrish@sms.ed.ac.uk`
  - Neofytos Themistokleous: `s1770805@sms.ed.ac.uk`
  - Estifa'a Zaid: `Estifa'a.Zaid@ed.ac.uk`
- Semester-2:
  - Joe Bradley: `s1406647@sms.ed.ac.uk`
  - Nisha Grewal: `N.Grewal@sms.ed.ac.uk`
  - Kevin Lo: `s2268228@sms.ed.ac.uk`
  - Patrick Sinclair: `s1212500@sms.ed.ac.uk`
  - Neofytos Themistokleous: `s1770805@sms.ed.ac.uk`
  - Tianyi Yang: `s2109053@sms.ed.ac.uk`

1.3 Pre-requisites

1. Undergraduate students who wish to take this course must have successfully taken
   - *Modern Physics* (`PHYS08045`)
   - *Fourier Analysis and Statistics* (`PHYS09055`) or *Probability* (`MATH08066`)
   - *Computer Modelling* (`PHYS09057`) or *Numerical Recipes* (`PHYS10090`)
2. Students must have programming experience and feel comfortable using
   - (Mandatory) `python`
   - (Desirable) `C++`, `NumPy`, `SciPy`, `Jupyter` notebooks

If any of these conditions is not satisfied, please seek permission from course organiser before enrolling on DAML.

It is also recommended (but not mandatory) that students enroll on *Detectors in Particle & Nuclear Physics* (`PGPH11104`).

1.4 Lectures and workshop sessions

Semester 1:

- Lectures: Monday 9-9:50 PM (JCMB LTB)
- Workshops: Wednesday 10 AM-1 PM (JCMB 4325D)
- Lectures start on Week-1: 20^{th} September

Semester 2:

- Lectures: Monday 10-10:50 AM (LTB)
- Workshops: Wednesday 10 AM - 1 PM (JCMB 3210)
- Lectures start on Week-1: 17^{th} January

1.5 Syllabus and schedule

Semester 1:

Week/CP	Topic	Lecturer
20 Sep: W1	Intro to Probabilities: Frequentist & Bayesian (†)	Christos
27 Sep: W2	Data Science tools & Decision Trees (∗)	Guillermo
04 Oct: W3/CP1	Intro to Neural Networks & Deep Learning	Guillermo
11 Oct: W4	Intro to `C++` (∗)	Ben
18 Oct: W5/CP2	Intro to Convoluted Neural Networks	Guillermo
25 Oct: W6/CP3	Intro to Generative & Adversarial Neural Networks	Guillermo
01 Nov: W7/CP4	Pseudo-random generators; Probability densities	Pete
08 Nov: W8	Fitting: χ^2, Maximum Likelihood (∗)	Pete
15 Nov: W9/CP5	Data selection, p-values, significance interpretation	Christos
22 Nov: W10/CP6	Fitting: parameter estimation	Pete
29 Nov: W11/CP7	Fit errors, correlated errors, systematic uncertainties	Pete
(†): No workshop or checkpoint this week (∗): Workshop taking place, but checkpoint not assessed		

Semester 2:

Week/CP	Topic	Lecturer
17 Jan: W1/CP8	Intro to GEANT & Simulated physics models	Ben
24 Jan: W2	Hypothesis tests: Type-I/II errors; Likelihood ratios (∗)	Christos
31 Jan: W3/CP9	GEANT & MC-truth	Ben
07 Feb: W4/CP10	Discoveries, Exclusion and Upper Limits	Christos
14 Feb: W5/CP11	GEANT, energy scales and detector interactions	Ben
Flexible learning week		
28 Feb: W6/CP12	Neural Network Regression (LHCb app)	Christos
07 Mar: W7	GEANT & Neural Networks (∗)	Ben
14 Mar: W8/CP13	Neural Network Classification (ATLAS app)	Christos
21 Mar: W9/CP14	Gradient Boosted Regression Trees (TITUS app)	Christos
28 Mar: W10	Averaging correlated measurements (∗)	Pete
(∗): Workshop taking place, but checkpoint not assessed		

1.6 Assessment

Data Analysis and Machine Learning is a continuously assessed course. The overall DAML assessment will consist of two parts:

- The sum of marks achieved while carrying out the weekly checkpoints in the workshops (14 in total) will count for 49% of the total course mark. The maximum

number of points for the checkpoint in a given workshop corresponds to 3.5% of the total course mark. All workshops have the same weight.

Checkpoints will be made available on the weekend before the Monday morning lecture. Students are expected to come prepared to the workshop and be able to work on the checkpoints right away. The deadline for handing in the answers to the checkpoints will be **at 10 AM on the first Friday after the workshop**, i.e. before the next checkpoint becomes available. There is one exception: the Semester-1, Week-11 CP7 (to be discussed on the Monday 29 November lecture) is due exceptionally **on Tue 18 January at 10 AM**.

If you miss a workshop (and the corresponding checkpoint), please get in touch with the course organiser to arrange for out-of-hours access to the lab and the completion of the missed checkpoints.

- The marks obtained for the DAML projects will count for 51% of the total course mark. There will be four projects in total, out of which the **best three** submitted and marked reports will contribute to your total mark (with the maximum grade of each project equal to 17% of the total course mark). The deadlines for handing in the project reports are as follows:
  - First report: **Fri 3 Dec at 4 PM**
  - Second report: **Fri 28 Jan at 10 AM**
  - Third report: **Fri 11 Mar at 10 AM**
  - Fourth report: **Fri 8 Apr at 4 PM**

1.6.1 Late submission penalty

The penalty for late submission is a reduction of the mark (of the CP or project) by 5% of the maximum obtainable mark per calendar day (e.g. a mark of 65% on the common marking scale on a CP would be reduced to 60% up to 24 hours later). This applies for up to seven calendar days, after which a mark of zero will be given. The original unreduced mark will be recorded by the School and the student informed of it.

1.7 Course material

Lecture notes and checkpoint assignments will be uploaded on Learn.

2 Bibliography

Main Books:

- *"Data Analysis in High Energy Physics"*, by Behnke, Kröninge, Schott, and Schörner-Sadenius

  - Print ISBN: 9783527410583 , Online ISBN: 9783527653416
  - DOI: 10.1002/9783527653416
  - Open access: `https://onlinelibrary.wiley.com/doi/book/10.1002/9783527653416`

- *"Statistical Methods for Data Analysis in Particle Physics"*, by Luca Lista
  - Print ISBN: 978-3-319-62839-4, Online ISBN: 978-3-319-62840-0
  - DOI: 10.1007/978-3-319-62840-0
  - Free access from UoE network

- *"Hands-On Machine Learning with Scikit-Learn and TensorFlow"*, by Aurélien Géron
  - ISBN: 978-1491962299

- *"An Introduction to Statistical Learning"*, by James, Witten, Hastie, and Tibshirani
  - Print ISBN: 978-1461471370, Online ISBN: 978-1-4614-7138-7
  - DOI: 10.1007/978-1-4614-7138-7
  - Open access: `https://web.stanford.edu/~hastie/ISLRv2_website.pdf`

- *"The Elements of Statistical Learning"*, by Hastie, Tibshirani, Friedman
  - Print ISBN: 978-0-387-84857-0, Online ISBN: 978-0-387-84858-7
  - DOI: 10.1007/978-0-387-84858-7
  - Open access: `https://web.stanford.edu/~hastie/ElemStatLearn/`

Computing:

- A large number of (free) online resources. In particular, we highly recommend the Codecademy `python` and `C++` courses (`https://www.codecademy.com/`).
- Local (and informal) tutorials on `python` and `C++` are typically organised, if there is sufficient interest.